Unboxing the Concepts in Newcomb's Paradox: Causation, Prediction, Decision in Causal Knowledge Patterns

Roland Poellinger*

In Nozick's rendition of the decision situation given in Newcomb's Paradox dominance and the principle of maximum expected utility recommend different strategies. While evidential decision theory (EDT) seems to be split over which principle to apply and how to interpret the principles in the first place, causal decision theory (CDT) seems to go for the solution recommended by dominance ("two-boxing"). As a reply to the CDT proposal by Wolfgang Spohn (2012), who opts for "one-boxing" by employing reflexive decision graphs, I will draw on the framework of causal knowledge patterns, i.e., Bayes net causal models (cf. e.g. Pearl 2009), augmented by non-causal knowledge (epistemic contours), to finally arrive at "one-boxing" – more intuitively and more closely to what actually is in Nozick's story. This proposal allows the careful re-examination of all relevant concepts in the original story – it prompts a re-evaluation of how *prediction* may be analyzed, philosophically and formally, and what the decision-maker's conceptualization of the situation might look like.

Keywords: evidential vs causal decision theory, Newcomb's paradox, Bayes nets, causal models, interventionist account of causation

---

Decision theory in general examines the rational principles guiding the decisions that aim at the attainment of one's goals. *Causal* decision theory (CDT) does so by taking one's act's consequences into account – rationally choosing an option must be based on the available knowledge about the causal relations in the respective situation, so the argument goes. One of the principles taken to be a measure for rationality is the option of maximizing the utility of the outcome, i. e., by making the outcome equal or better than if one had chosen a different alternative for action. Probabilities and utilities are used to compute an act's expected utility such that – as emphasized in *causal* decision theory – dependence between acts and outcomes

*Munich Center for Mathematical Philosophy (LMU Munich), r.poellinger@lmu.de

are understood as of causal (asymmetrical) character – contrary to a merely *evidential* theory of decision making (EDT). A second principle of rationality dictates choosing the course of action that is better, regardless of what the world is like. This principle of dominance seems to be in conflict with the first-mentioned principle of expected-utility maximization in the curious case of Newcomb's paradox.

Newcomb, Nozick, and a problem

Referring back to the physicist William Newcomb, who first formulated this dilemma for decision theory, Robert Nozick elaborates on – as he calls it – Newcomb's problem, in which two principles of rational choice seemingly conflict each other, at least in numerous renditions in the vast literature on this topic.[1]

In Newcomb's problem some human-like agent plays a game against some daemon predictor that influences the course of the game upon predicting his opponent's move. The agent may choose to take either one or two boxes in front of him – either box 1 only or box 1 and 2 together. In doing so he has no knowledge about the contents of the opaque box 1, but he can see one thousand dollars placed in box 2. If the daemon predicts that the agent will take only one box (i. e., box 1), he will put one million dollars in the opaque box 1. The daemon will put nothing in box 1, though, if he foresees the agent taking both boxes. The prediction is reliable, or as Nozick introduces the predictor, "[o]ne might tell a longer story, but all this leads you to believe that almost certainly this being's prediction about [the agent's] choice in the situation to be discussed will be correct."[2] Moreover, the agent has perfect knowledge of all these features of the decision game he finds himself in.[3]

The possible outcomes of the game are presented in table 1 where the rows stand for the agent's options, the columns partition the world in possible states, and each cell contains the sum our agent receives upon choosing an action in some state of the world.

[1] Cf. (Nozick, 1969) for the original presentation of the paradox and (Weirich, 2008) for an overview on various suggestions of how to solve the Newcomb case.

[2] Cf. (Nozick, 1969, p. 114).

[3] Note that for reasons of simplicity this presentation of the Newcomb game situation slightly (yet inessentially) differs from the way Nozick originally presents it in (Nozick, 1969).

	prediction: one-boxing	prediction: two-boxing
take box 1	\$ 1M	\$ 0
take box 1 and 2	\$ 1M + \$ 1T	\$ 1T

Table 1: Possible outcomes in Newcomb's problem for the options of taking box 1 only, taking boxes 1 and 2, respectively, and for correct and incorrect predictions made by the daemon.

Now, what makes Newcomb's case so problematic is the fact that the choice of action seems to depend on the choice of the principle one applies in *rationalizing* the situation. Two principles seem to be concurring candidates in reasoning about Newcomb's problem, which – although unrealistic – seems to trigger solid intuitions about the decision-theoretic norms to be applied here.[4] The rationales of maximizing expected utility and of choosing dominating options are defined in the following.[5]

Definition 0.1 (**Maximum Expected Utility**)
Among those actions available, one should perform an action with maximal expected utility*.*
The expected utility $EU(A)$ of an action A yielding the exclusive outcomes $O_1, \ldots, O_n$ with probabilities $P(O_1), \ldots, P(O_n)$ and corresponding utilities $U(O_1), \ldots, U(O_n)$ is calculated by the weighted sum

$$\sum_{i=1}^{n} P(O_i) \times U(O_i).$$

Definition 0.2 (**Dominance**)
If there is a partition of world states such that, relative to it, action A weakly dominates action B, then A should be performed rather than B.
Action A weakly dominates *action B for person P iff, for each state of the world, P either prefers the consequence of A to the consequence of B, or is indifferent between the two consequences, and for some state of the world, P prefers the consequence of A to the consequence of B.*

[4] Nozick himself obviously put the story on the test bench: "I should add that I have put this problem to a large number of people, both friends and students in class. To almost everyone it is perfectly clear and obvious what should be done. The difficulty is that theses people seem to divide almost evenly on the problem, with large numbers thinking that the opposing half is just being silly." – cf. (Nozick, 1969, p. 117).

[5] The following definitions are adapted from (Nozick, 1969, p. 118).

Let us take 'reliable' (as ascribed to the daemon's faculty of foreseeing future events) at face value and compute the expected utility for the outcome of each specific course of the game – the unit of the expected utility being dollars in our case. Assuming a reliable daemon basically amounts to saying that the act of taking one or both boxes and the prediction of this very act are correlated in a way such that acts in states of the world with incorrect predictions receive a probability of 0, whereas matching acts and predictions receive the probability of 1. Table 2 shows the expected utilities for all four thinkable courses of the game with one option clearly to be preferred over all others: The agent should take only the opaque box and can then be certain of winning $ 1M, which clearly supercedes the alternatives as *maximum expected utility*.

	prediction: one-boxing	prediction: two-boxing
take box 1	$ 1M	$ 0
take box 1 and 2	$ 0	$ 1T

Table 2: Computing expected utilities in the case of a perfectly reliable prediction yields the utility of $ 0 for all cells representing incorrect predictions. Maximizing this expected utility amounts to choosing only box 1.

Pondering a different approach to maximizing the outcome of the game, NOZICK tweaks the story a little: The predictor did make his prediction a week ago, and it is now the agent's turn to make up his mind and take either only the opaque box 1 or on top of that also the transparent second box 2, which contains one thousand dollars openly visible to the agent. The money is already there and will not be taken out of the boxes anymore after the agent has made a decision. So, regardless of the daemon's prediction, adopting the *principle of dominance* forces the agent to take both boxes – he will always end up with one thousand dollars more than if he had only taken one box. Taking both boxes even *strictly dominates* the act of taking only one box as can be read off table 1 by comparing an entry in the second line to the entry in the first line within the same partition of the world's states.

Obviously, the principle of maximizing expected utilities and the principle of dominance yield opposing recommendations to the deliberating agent. While standard *evidential* decision theory seems to lean towards *one-boxing* (taking an agent's act as a sign of what the prediction must have been), *causal* decision theorists clearly position themselves on the side of *two-boxing*,

rejecting backward causation and understanding the agent's deliberate decision as cutting any connection between act and prediction. This very idea of cutting links by deliberately producing specific events has been made mathematically precise in network models, e. g., as in (Spirtes et al., 2000), and is most elaborately presented in Judea PEARL's book *Causality* (2000): The concept of causally efficacious control is formally understood as an *external intervention* on a certain node in the network, represented as a local surgery in the graph, and structurally expressed as a transformation of the model. Causal asymmetry can be mimicked this way. When PEARL, within this *interventionist* account of causal reasoning, discusses *model-internal* observed acts and *model-altering* actions from outside, he also comes to reflect upon the conceptual difficulties hidden in Newcomb's problem:

> *The confusion between actions and acts has led to Newcomb's paradox (Nozick 1969) and other oddities in the so-called evidential decision theory, which encourages decision makers to take into consideration the evidence that an action would provide, if enacted. This bizarre theory seems to have loomed from Jeffrey's influential book* The Logic of Decision *(Jeffrey 1965), in which actions are treated as ordinary events (rather than interventions) and, accordingly, the effects of actions are obtained through conditionalization rather than through a mechanism-modifying operation like* $do(x)$.[6]

When PEARL goes on by comparing the maxims of evidential and causal decision theory, he baldly comments in a footnote:

> *I purposely avoid the common title "causal decision theory" in order to suppress even the slightest hint that any alternative, noncausal theory can be used to guide decisions.*[7]

To reconcile the dominance principle with the expected-utility principle – and hence to dissolve the paradox in Newcomb's case – has been the aim of quite a few proposals, which nevertheless arrive at different conclusions.

Conditionals and causal graphs

In *A Theory of Conditionals* (1968) Robert STALNAKER suggests a formal framework for analyzing the truth of counterfactual statements: 'If A, then

[6] Cf. (Pearl, 2009, p. 108). PEARL's $do(\cdot)$-operator precisely does the job of setting a variable X to a constant value x, thereby deactivating any causally described relation between this variable and its parents in the structure.

[7] Cf. (Pearl, 2009, p. 108, footnote 1).

B' is assigned a truth value in accordance with the following informal condition:

> *Consider a possible world in which A is true, and which otherwise differs minimally from the actual world.* 'If A, then B' is true (false) just in case B is true (false) in that possible world.[8]

The subjunctive connective '$>$' is subsequently equipped with the more formal semantical rules

$$A > B \text{ is true in } \alpha \text{ if } B \text{ is true in } f(A, \alpha) \text{ and}$$
$$A > B \text{ is false in } \alpha \text{ if } B \text{ is false in } f(A, \alpha),$$

where α is a possible world, the *base world*, and $\beta = f(A, \alpha)$ represents the *selected world* minimally differing from the actual world in which B is evaluated (with f being the selection function operating on a suitable similarity ordering of possible worlds).

Now, in his *Letter to David Lewis* (1972) Stalnaker suggests a way of calculating expected utilities in the Newcomb problem that uses probabilities of counterfactual conditionals instead of standard conditional probabilities.[9] The expected utility of some action A would then be computed the following way:

$$EU(A) = \sum_{i=1}^{n} P(A > S_i) \times U(A \,\&\, S_i),$$

where n indicates the number of states S the world is partitioned into, i. e., $n = 2$ for the two possible predictions 'one-boxing' ($i = 1$) and 'two-boxing' ($i = 2$). As Stalnaker argues, the agent's action does not *cause* the daemon's prediction made in the past, and hence the probability of the conditional equals the probability of the prediction alone. But this sets all probability terms in the sum formula above to equal values – the utilities can just be read off the corresponding cells in table 1. Two-boxing's expected utility will always be greater than one-boxing's expected utility. Following Robert Stalnaker's suggestion of interpreting the involved probabilities *causally*, the maximization of expected utility and the dominance principle recommend taking the same action: two-boxing.

[8] Cf. (Stalnaker, 1968, p. 169).

[9] Cf. for this and the following (Weirich, 2008, sect. 2.2).

Applying causal decision theory to Newcomb's problem has been criticized by many authors – mainly because it yields the recommendation of taking both boxes, oftentimes dubbed 'counter-intuitive', which nevertheless remains as the only rationally explained choice given the circumstances of Newcomb's problem with decisions screening off acts from any previous events, as causal decision theorists claim. In his seminal book *The Foundations of Causal Decision Theory* James Joyce clearly states his position on the issue:

> *When the evidential and the causal import of actions diverge [...], the evidential theory tells decision makers to put the pursuit of good news ahead of the pursuit of good results. Many philosophers, I among them, see this as a mistake. Rational agents choose acts on the basis of their* causal efficacy*, not their auspiciousness; they act to* bring about *good results even when doing so might betoken bad news.*[10]

While, e. g., David Lewis and Brian Skyrms in their accounts mark attainable situations by building causal information into states of the world and thereby reconcile the above otherwise diverging principles of rational choice in the recommendation of two-boxing, Ellery Eells in his considerations arrives at the same conclusion without drawing on the notion of causality. He claims that mere reflection on the available *evidence* will force the agent to rationally go for both boxes – even more direct without the recourse to any causal theory. Quite in this line of reasoning Richard Jeffrey also eliminates any hint of a causal nexus between the events in Newcomb's problem for the sake of a less metaphysically charged analysis. Pondering the Newcomb case Jeffrey seems to oscillate between one-boxing and two-boxing to later arrive at the conclusion that the story, presented this way, is a somehow *illegitimate* decision problem with the freely deliberating agent not capable of freeing his decision from being correlated with the predictor's prediction.[11] Terry Horgan and Paul Horwich take the Newcomb plot at face value and promote one-boxing, simply because one-boxers ultimately take more money home, as the story is told. Paul Weirich diagnoses dryly: "The main rationale for one-boxing is that one-boxers fare better than do two-boxers. Causal decision theorists respond that Newcomb's problem is an unusual case that rewards irrationality. One-boxing is irrational even if one-boxers prosper."[12]

[10] Cf. (Joyce, 1999, p. 146).
[11] Cf. e. g. (Joyce, 2007).
[12] Cf. (Weirich, 2008, sect. 2.5).

Having developed his ranking theory as a tool for epistemology and causal analysis,[13] Wolfgang Spohn positions himself on the side of causal (vs. evidential) decision theory and had been a strong advocate of two-boxing for a long time before he started "Reversing 30 Years of Discussion" by presenting an elaborate argumentation "Why Causal Decision Theorists Should One-Box."[14] Spohn's primal commitment can be found in the title of his paper *Bayesian nets are all there is to causal dependence* (2000). In such Bayes net causal models (generic) events are encoded as random variables and graphically represented by single nodes. A node X is connected to its parents by a set of directed edges, which jointly mark the causal mechanism responsible for bringing about some specific value x (of X). In Judea Pearl's framework these mechanisms are defined as deterministic functions potentially also taking some disturbance variable as an argument to represent (observational) uncertainty in the model. The Markov compatibility of the graph and the corresponding probabilistic model can in causal terms be interpreted as *causal Markov condition*: Causes screen off their direct effects particularly from prior influences and more generally from changes in any other event that is represented as a non-descendant in the graphical rendition (when the graph encodes precisely the perceived causal independencies of the modeled situation).

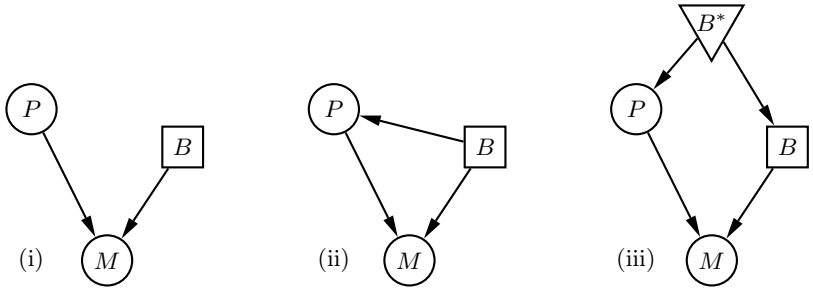

Figure 1: Wolfgang Spohn discusses the usual *manipulated (mutilated) causal graph* (i) employed by causal decision theorists for the analysis of the Newcomb problem, the *decision graph* (ii) for the same situation, and the *reflexive decision graph* (iii) augmented by the decision node B^*.

[13] Cf. (Spohn, 2012a).

[14] The quotations here are taken from the title of Spohn (2012b).

Now, figure 1 illustrates the golden thread in SPOHN's chain of reasoning in the Newcomb case. Time evolves from top to bottom in all three Bayes net diagrams. The left diagram (i) shows the standard rendition used by causal decision theorists for the analysis of the paradox – this *mutilated causal graph* contains the node P representing the daemon's prediction as the first event in time before *action* node B (representing the agent taking one or two boxes) and the bottom node M (for monetary outcome). The diagram is *mutilated* quite in agreement with PEARL's interventionist framework: The hypothetical local surgery, i. e., the deliberate intervention on B, prunes any arrows possibly pointing towards B, thereby freeing this node from the influence of any other node in the model and making the corresponding variable an exogenous one. The course of action can now be chosen on the basis of this *decision graph*, in which the *wiggled* variable is graphically represented by the square node. This rendition follows the two decision-theoretic principles highlighted by SPOHN in this context: "acts are exogenous" and – derived from the first – "no probabilities for acts." Of course, SPOHN's *acts* have to be interpreted as PEARL's *actions* (i. e., *acts* in *mutilated* models). Whatever the connection between nodes P and B might have been in some graphical rendition of the original causal relations understood as representing the Newcomb plot (e. g., with P as a direct cause of B), graph (i) in figure 1 represents the variables' dependencies once the agent deliberately takes action. In Bayes net terms P and B are d-separated (by the collider in $P \rightarrow M \leftarrow B$), which makes the choice of taking both boxes rational – whatever has been put into the boxes (based upon the prediction early in the game) will not become less by choosing either one or, alternatively, two boxes (later in the game). SPOHN declares himself dissatisfied with this analysis and brings up the mind-bugging questions about the reliability of the daemon, again:

> *What about the remarkable success of the predictor that suggests that given you one-box it is very likely that she will have predicted that you will one-box, and likewise for two-boxing? How do [these considerations] enter the picture? They don't. [Causal decision theorists] do not deny them, but they take great pains to explain that they are not the ones to be used in practical deliberation calculating expected utilities; and they diverge in how exactly to conceive of the subjective probabilities to be used instead.*[15]

If the causal graph contained one more arrow from B to P, making the agent's action a direct cause of the daemon's prediction (as illustrated in

[15] Cf. (Spohn, 2012b, p. 4).

figure 1, diagram (ii)), we would inevitably introduce backward causation into the analysis. SPOHN wants to avoid this but interprets graph (ii) as the *decision-guiding pattern* which the agent uses to choose between alternative actions – in SPOHN's terms: the ordinary decision graph for Newcomb's problem. How are the causal relations laid out, however? If neither the prediction causes the agent's act nor this act causes the daemon's prediction, we have to infer the existence of an earlier third event as a common cause of both P and B – quite in accordance with REICHENBACH's *Common Cause Principle*. SPOHN's straightforward suggestion is to understand the *decision situation* the agent finds himself in as the common cause in question. This decision situation B^* (as introduced into graph (iii) in figure 1) might consist of all the agent's beliefs, prior knowledge, or rational principles the agent may not even be aware of (the daemon is, however) but which he will without fail employ in deciding about his strategy B when standing before the two boxes. In particular, B^* also contains the full ordinary decision-guiding pattern (ii), which makes graph (iii) a *reflexive decision graph* containing a reduced version of itself.[16] Making this move, SPOHN openly rejects the "acts are exogenous" principle. An agent's strategic deliberation about alternative courses of action does *not* decouple the act from past or future events – he might, quite on the contrary, make his deliberations depend on (i. e., graphically speaking, link them to) *predecessor nodes in the diagram*. He might, on top of that, also be aware of the probabilities of different actions he may choose from, knowing what he *usually does* or intentionally *avoids in normal cases* etc. There might be *probabilities for the agent's act*, after all. Querying SPOHN's reflexive decision graph on the ground of all these considerations ultimately yields the recommendation of one-boxing – after reflecting on the current situation (in B^*), the rational agent must come to the unequivocal conclusion that deciding to one-box and acting accordingly simply maximizes the utility of his act B.

Let us compare SPOHN's analysis with PEARL's causal maxims, once more. The ordinary decision graph (as displayed in figure 1.ii) fully complies with what PEARL would devise for strategic reasoning, i. e., a graph that simulates possible outcomes of hypothetical interventions. Setting B tells us the value of M. B is an exogenous variable such that the "acts are exogenous" principle is adhered to – act and action amount to the same consequence in this case. The evidential and the causal approach perfectly concord in

[16] SPOHN gives clear rules for the step-wise reduction of a reflexive decision graph to its ordinary counterpart possibly containing backward links – cf. (Spohn, 2012b, sect. 3).

this diagram, were it not for the directed backward edge $B \to P$. This is the reason for the causal decision theorist to think directly in terms of the mutilated graph (given in figure 1.i) and for SPOHN to call diagram 1.ii not *causal* but *reduced, ordinary decision graph*. In the further step of construing the reflexive decision graph 1.iii, SPOHN must reject the "acts are exogenous" principle and convincingly argues for his case: The hypothetical intervention on the variable B must not be performed within the reflexive decision graph. This graph makes explicit what it means for the agent to be rational, i. e., he acts on his knowledge, principles, and rational considerations given in B^*. Pruning the link $B^* \to B$ would make the agent plainly *irrational* and *ignorant* of his own situation, since the deliberation process is *pushed into the model*, after all.

Technical answers to questions about how to properly reduce reflexive decision graphs to their ordinary, structural counterparts can all be found in SPOHN's explications. Conceptual questions remain, however. Firstly, the introduction of a common cause for B and P essentially adds to the Newcomb story the idea of being (perhaps physically determinately) *pre-disposed*. In a way, this metaphysically overloads the already artificially construed plot with another element just by drawing on REICHENBACH's principle of the common cause. Moreover, it forces SPOHN to set apart the agent's inclinations to take certain actions from the acts themselves. Decision making in the game is consequently *re-interpreted as only discovering one's previously fixed inclinations* (where discovery is not something brought about actively, e. g., such that it would manifest itself in deliberate, hypothetical test interventions, but simply a feature of persistent rationality becoming *evident*). This rendition seems very far from the much more intuitive interventionist framework, which merely requires the agent to bear a *confined mini laboratory* in his head and turn the knobs therein – knowledge about the mechanisms will yield unique virtual outcomes and guide decision making. Nevertheless, SPOHN's complex reflexive decision graph does rest in its core on the very simple ordinary reduced decision graph (figure 1.ii) to which the whole burden of explanation is shifted. This shall be looked at more closely in the following. What can be the content of this reduced graph, after all? If the link $B \to P$ is dismissed as a causal relation, of what nature can it be? If it, on the other hand, does stand for some hidden causal connection and is dismissed as an instance of backward causation, it must represent a causal link through some obscure common cause. If this common parent node of both P and B is the decision situation again – just as in the reflexive graph

on the meta level – analysis enters an infinite regress at this point. Only the interventionist approach could prevent this from happening by pruning $B \rightarrow P$, but then this would already apply on the upper level in the reflexive decision graph and conflict with SPOHN's final conclusion. If the supposed common cause in figure 1.ii is interpreted as some irreducible obscure past event or state whose existence just has to be acknowledged and whose link to B shall not be interrupted, then how would it be possible to perform hypothetical test interventions on this very node to virtually maximize the outcome? If reflecting on this graph ultimately comes down to just *observing* the propagation of values, then, one has to conclude, SPOHN's suggestion is constrained to stay within evidential reasoning.

The concepts involved

As in all cases of paradoxical disagreement, the concepts involved should be explicated as precisely as possible to do away with any source of confusion and to fix the premises prior to systematic treatment. In the Newcomb case *causation*, *decision*, and *prediction* take center stage – a closer look at these concepts is in order. First, causal relations, taken to be directed in accordance with time, shall be understood as encodable and storable in Bayes net causal models as devised by Spirtes et al. (2000) or Pearl (2009). These models structure stable and deterministic dependencies as perceived by the modeler or given in the data (once observational noise is recognized as such). PEARL's causal models compactly represent a set of counterfactual situations by allowing for hypothetical local interventions on certain variables in the structure – as structured bodies of knowledge they can be used for communicative purposes and facilitate explanation, instruction, and prediction. The decision maker takes all available data into account, he has knowledge of all causally relevant information, i. e., of the full model with all dependencies. Hypothetical interventions on action variables in the model will yield predictions about potential outcomes such that decision making is finally guided by computing and optimizing outcome values (in a maximum utility approach). The prediction (computation) of outcomes by the pondering agent is to be distinguished from the kind of prediction the daemon performs as a move *within* the game situation. In accordance with intuition and our use of language, the daemon's miraculous faculty has to do with knowing or learning things by *seeing into the future*. This concept of prediction seems bi-directional, as the content of the prediction and the predicted event stand

in close relation. Knowing one makes the other inferable.[17] It is worth noting that the accuracy (i. e., the degree of reliability) of the daemon's prediction does not decide between the evidential and the causal approach: CDT argues for 2-boxing on the basis of independence assumptions. And EDT will prefer 1-boxing over 2-boxing down to the low predictive accuracy of .5005 – close to exchanging prediction for a coin toss. The paradox is much rather about the principles that guide decision making, the use of language, and our intuitions about the concepts involved.[18] Now, if we actually wanted to base our strategy on causal knowledge, we arrive at the central question of this paper: If it does neither seem right to say that the prediction causes the predicted event nor that the predicted event causes the prediction (through some backward connection), how could the concept of prediction (as part of the game situation) be suitably accommodated in a causal model to guide decision making?

Integrating causal and non-causal knowledge

What is needed for the integration of a prediction link into standard causal models is the introduction of a new type of edge – a non-directed, non-causal but rather informational link, capable of propagating information instantaneously, and moreover not to be deactivated by any means. This link should work like synonyms, mathematical inter-definitions, or logical relations (which certainly all belong to the pool of knowledge we use for decision making). In standard statistical modeling prediction and predicted event would be collapsed into one single variable (node, respectively). In philosophical context we would like to disambiguate conceptually, and in the causal model the temporal distance between prediction and predicted event should find its expression. Consequently, the final model ought to contain two distinct nodes and mark these nodes as tightly, functionally dependent.

[17] In this characterization prediction works much like a quotation relation.

[18] Some people (even advocates of causal decision theory) argue that it makes a significant difference if the story is told in deterministic terms (with a fully reliable predictor) or with indeterministically inaccurate predictions. Nozick (1969) comments critically:

> [Do these people] really wish to argue that if [they know] the prediction will be correct, [they] will take only the second, but that if [they know the prediction] will be wrong once in every 20 billion cases, [they] will take what is in both boxes? Could the difference between one in n, and none in n, for arbitrarily large finite n, make this difference?

When Judea Pearl writes about the principles of encoding causal dependencies in formal models, he notes:[19]

> *The ability to represent functional dependencies would be a powerful extension from the point of view of the designer. These dependencies may easily be represented by the introduction of deterministic nodes which would correspond to the deterministic variables. Graphs which contain deterministic nodes represent more information than d-separation is able to extract; but a simple extension of d-separation, called D-separation, is both sound and complete with respect to the input list under both probabilistic inference and graphoid inference.*

Along these lines and in addition to the directed edges representing factors of causal mechanisms in the graph, we straightforwardly introduce a type of non-directional, informational link, which shall be called *Epistemic Contour (EC)* to underline its intensional nature. For the purpose of this paper we shall restrict ourselves to introducing a single EC, described by a 1-1 function, that will be tested only by atomic interventions in the model. Integrating such an epistemic contour in the causal model turns this into a model of hybrid knowledge, i. e., a rich structure of directed and undirected relations, in the following referred to as *Causal Knowledge Pattern (CKP)*.[20]

Prediction is a matter of knowledge

What the backward link $B \rightarrow P$ in graph 1.ii can possible mean shall in the following be made explicit within a very simple causal knowledge pattern, thereby ideally revealing more about the nature of the paradox and hopefully illuminating some more features of how we reason with (non-)causal knowledge. The CKP in figure 2 traces the story of Newcomb's problem by only referring to the events that actually are in the narration. The problem is not treated by tweaking the story but by choosing a framework fit to accommodate all relevant concepts.

[19] Cf. Verma and Pearl (1988), where d-separation is introduced as a means to discover independencies in the graph.

[20] Since context always disambiguates whether *'epistemic contour'* or *'EC'* refers to the functional description or its graphical representation, I will use the term for either. One technical remark is in order though: The alert reader will have noticed that the introduction of an undirected EC in DAG models will in general render those cyclic and non-Markovian. For the present case this does not pose problems – inference from the simple CKP proposed below will be computationally straightforward. To make CKPs in general useful for consistent computation, I elsewhere lay out principles of design and inference for a CKP framework.

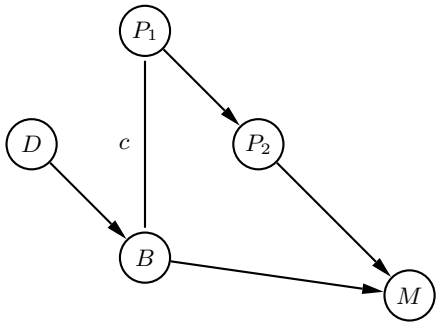

Figure 2: Newcomb's problem with the act of taking one or two boxes (B) tightly connected to the daemon's reliable prediction ($P1$) by an epistemic contour (c) in this causal knowledge pattern.

Our human-like agent deliberates about the situation he finds himself in and decides what to do (D), namely if he takes one box or both boxes (B). The daemon predicts what the agent will do (P_1) and prepares the boxes accordingly (P_2). The monetary outcome (M) should finally reward the rational agent. Time evolves from top to bottom in the diagram. The vertical positioning of P_2 is inessential for the analysis of the situation (P_2 could as well come after B if the game is set up in a way that the agent only writes down his choice on a sheet of paper secretly in step B). The daemon's prediction together with its reliability is interpreted in this causal knowledge pattern as an undirected 1-1 relationship. Neither would we say that the agent's act genuinely causes the prediction of this very act, nor does it sound right to say the prediction causes the predicted event.[21] But there is more in the pattern: B is not directly linked to the daemon's preparation of the boxes P_2 – this connection is mediated by the prediction P_1, which has direct causal influence on P_2 in turn. This is quite in agreement with Spohn's analysis that the causal structure of the Newcomb problem should exhibit some node prior to both players' acts in the game which at the same time takes care of the bidirectional transfer of information. P_1 and P_2 are separated in the proposed causal knowledge pattern for this very reason. On the other side, D (the human-like agent's decision situation) and B (his concrete move in the game – either taking one or both boxes) are separated, as well, to disentangle conceptually what it means for the agent to spontaneously and possibly

[21] Moreover, as is argued here, drawing on Reichenbach's *Common Cause Principle* for an explication of *prediction* is precisely a source of counter-intuitive inference.

unforeseenly *change his mind*. This is a much-discussed issue in the literature and does pose additional problems if the modeling allows for the agent changing his mind and the daemon's prediction referring to the 'wrong' decision. Not so in the suggested causal knowledge pattern, which links the prediction P_1 to the agent's final act B however often he may have made up or changed his mind before actually taking only one or, after all, both boxes. In other words, pondering courses of action must focus on B bearing the whole burden of explanation in the process of finding the best strategy for the maximization of the outcome. This is precisely how NOZICK tells the story.

The modeling does not draw on the insertion of backward links that would indicate backward causal flow. Nevertheless, *information* is transferred back in time along the epistemic contour c, thereby formally grasping the very meaning of 'prediction'. c will not get cut off by any local surgery in the graph. By suitably applying hypothetical test interventions the following claims can be read off the causal knowledge pattern – quite in accordance with intuition:

- The agent's decision (D) causes his act (B) – in general: any causal history of (B) naturally influences the agent's act causally;
- the agent's decision (D) is also interpreted as causing the daemon's peculiar prediction (P_1) and thereby also as causing the daemon's particular move in the game (P_2);
- intuition also conforms with the claim that the agent's taking one or two boxes (B) causes his antagonist's preparation of the boxes – the predictor *reacts* to (B), after all;
- nevertheless, the agent's act (B) does *not cause* its own peculiar prediction (P_1) but *determines it uniquely* and – looking at the pattern from above – *simultaneously* though *backwards through time*.

Now, especially the last point reveals the core of the paradox and locates the difficulties in reasoning about the causal relations involved. Any attempt at solving the artificial plot of Newcomb's problem hinges on the question of how to embed the concept of reliably predicting future events into the formal analysis (if such an analysis is not denied in the first place exactly because of the fictional character of the narration). The causal knowledge pattern above presents the prediction as the very thing it is – an *image* of the agent's act. *Backward links* are excluded from this rendition while querying the pattern does yield *indirect causal claims referring back across time*.

This interpretation would of course not stand physically ontologically based scrutiny, but it conforms with our concepts of prediction (of future events) and reaction (to facts just learned of). How pieces of knowledge are organized and information propagated is shown in the causal knowledge pattern devised here. Obviously, the "acts are exogenous" principle insisted on by proponents of an interventionist account of causation is relativized in applying causal knowledge patterns to problems of decision theory. The epistemic contour c is not deactivated by intervening on B, while the one directed edge $D \rightarrow B$ is removed by the external action $do(B = b)$ – quite in PEARL's sense B and P_1 become "jointly exogenous". To sort the terms involved here: The *act* B becomes *exogenous* by virtue of the *action* $do(B = b)$, which is itself *external*.[22] If the prediction of events is formalized within a model (a causal knowledge pattern, respectively), *foreseeing acts* can be made explicit, while *foreseeing actions* cannot be given graphical expression. Reflecting on the Newcomb situation and performing hypothetical manipulations on the basis of integrating causal and non-causal knowledge finally guides the agent (who is aware of the setting) towards the correct decision. Resorting to reflexiveness is not necessary for virtually maximizing the outcome. The conclusion must be one-boxing.

As a concluding remark, David LEWIS shall be mentioned here once more. He examines another paradoxical puzzle of strategic thinking and finds in 1979 that the "Prisoners' Dilemma Is a Newcomb Problem", too.[23] The story in this particular dilemma shall be outlined briefly. Two suspects are caught by the police, that do not have sufficient evidence for conviction and therefore question the prisoners separately and (also separately) promise immediate release if the prisoners betray the respective other prisoner by confessing. However, if both confess, each serves a sentence of three months – in case both remain silent, each serves one month. Table 3 summarizes the situation compactly. If prisoner A applied the *principle of dominance* to his situation, he would of course confess, thereby always being off better than if he remained silent. If both prisoners think alike in this respect, however, they will be doomed to a sentence of another three months in prison. This is what makes the situation a strategic dilemma: Attributing the same (degree of) rationality to both prisoners does not entail the best outcome. If they include in their deliberations the ascription of *like-mindedness* to their fel-

[22] For clarification: *exogenous* remains a model-internal property of nodes (i. e., variables, respectively), whereas *external* marks transformations of causal structures.

[23] The quotation refers to the title of Lewis (1979).

low inmate, both of them should remain silent. If this ascription is reliable enough (or even deterministically certain), e. g., because of some commitment to the same gang code, then the prediction in Newcomb's problem and this theoretical simulation (the ascription) in the prisoners' dilemma essentially amount to the same thing – "[i]nessential trappings aside, Prisoners' Dilemma is a version of Newcomb's Problem, *quod erat demonstrandum*."[24]

	B stays silent	B confesses
A stays silent	Each serves 1 m	A serves 1 y, B goes free
A confesses	A goes free, B serves 1 y	Each serves 3 m

Table 3: Each of the prisoners could go free or serve a sentence of one month, three months, or a year – depending on their strategic decisions.

A *common* causal knowledge pattern might be used to capture all (non-) causal relations as in the above rendition of Newcomb's problem – quite naturally and without introducing further metaphysical assumptions about possible background variables. In fact, tilting the time axis in figure 2 by 90 degrees (such that time evolves from left to right) yields the skeleton of the prisoners' plot (of course, D and P_2 are particular ingredients of Newcomb's problem and inessential for the current examination). c is fit to represent mutual ascription of like-mindedness by both prisoners, who *must* decide to cooperate during their simultaneous (but separate) questioning to achieve the joint best result. May the Newcomb case be some fictional construction, Lewis makes the case for analyzing the prediction of future events and the ascription of like-mindedness to one's antagonist in terms of the same underlying pattern:

> *Some have fended off the lessons of Newcomb's Problem by saying: "Let us not have, or let us not rely on, any intuitions about what is rational in goofball cases so unlike the decision problems of real life." But Prisoners' Dilemmas are deplorably common in real life. They are the most down-to-earth versions of Newcomb's Problem now available.*[25]

[24] Cf. (Lewis, 1979, p. 239).

[25] This final quotation borrows the concluding paragraph from (Lewis, 1979, p. 240). I agree with Lewis on the point that situations of strategic deliberations of the kind exemplified here are "the most down-to-earth versions of Newcomb's Problem" – because there is *nothing more to know* than already said – in contrast to cases of so-called *medical Newcomb problems* where research might in most cases yield additional information and knowledge about true common causes whose influence would indeed be rendered void by free deliberation/active intervention.

References

Joyce, J. M. (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press.

Joyce, J. M. (2007). Are newcomb problems really decisions? *Synthese*, 156(3):537–562.

Lewis, D. (1979). Prisoners' dilemma is a newcomb problem. *Philosophy & Public Affairs*, 8(3):235–240.

Nozick, R. (1969). Newcomb's problem and two principles of choice. In Rescher, N., editor, *Essays in Honor of Carl G. Hempel*, pages 114–146. Dordrecht: Reidel.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. Adaptive Computation and Machine Learning. MIT Press.

Spohn, W. (2000). Bayesian nets are all there is to causal dependence. In Galavotti, M. C. et al., editors, *Stochastic Dependence and Causality*, pages 157–172. CSLI Publications, Stanford.

Spohn, W. (2012a). *The Laws of Belief: Ranking Theory and its Philosophical Applications*. Oxford University Press.

Spohn, W. (2012b). Reversing 30 years of discussion: Why causal decision theorists should one-box. *Synthese*, 187(1):95–122.

Stalnaker, R. C. (1968). A theory of conditionals. In Sosa, E., editor, *Causation and Conditionals (Readings in Philosophy)*, chapter XII, pages 165–179.

Verma, T. and Pearl, J. (1988). Causal networks: Semantics and expressiveness. In *Proceedings of the 4th Annual Conference on Uncertainty in Artificial Intelligence (UAI-88)*. Elsevier Science, New York.

Weirich, P. (2008). Causal decision theory. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, CSLI, Stanford University.